Department of the Interior
U.S. Geological Survey

Provisional Burned Area Essential Climate Variable (BAECV) Algorithm Description

By Todd Hawbaker[1], Susan Stitt[1*], Yen-Ju Beal[1], Gail Schmidt[2], Jeff Falgout[3], Brad Williams[3], and Josh Takacs[1]

1. U.S. Geological Survey, Geosciences and Environmental Change Science Center, Denver, CO
2. Stinger Gaffarian Technologies (SGT), contractor to the U.S. Geological Survey, Earth Resources and Observation and Science Data Center, Sioux Falls, SD
3. U.S. Geological Survey, Core Sciences, Analytics, Synthesis, and Libraries, Denver, CO

* Retired

Monday, March 16, 2015

Contents

Figures

Tables

Abbreviations and Acronyms

AUC	Area Under the Curve
BAECV	Burned Area Essential Climate Variable
CDR	Climate Data Record
CONUS	Conterminous United States
ECV	Essential Climate Variable
ESPA	Earth Science Processing Architecture
ETM+	Enhanced Thematic Mapper Plus
GBRM	Gradient Boosted Regression Models
LEDAPS	Landsat Ecosystem Disturbance Adaptive Processing System
MTBS	Monitoring Trends in Burn Severity
NBR	Normalized Burn Ratio
NDMI	Normalized Difference Moisture Index
NDVI	Normalized Difference Vegetation Index
QA	Quality Assurance
RMSE	Root Mean Squared Error
ROC	Receiver-Operator Characteristic
SLC	Scan Line Corrector
TM	Thematic Mapper
USGS	U.S. Geological Survey
WRS2	World Reference System, version 2

Abstract

The U.S. Geological Survey (USGS) has developed and implemented an automated algorithm to identify burned areas from Landsat scenes, producing Burned Area Essential Climate Variable (BAECV) products. These products include per-scene outputs of (1) the probability that a pixel was burned, given what was visible in the Landsat scenes, and (2) a burn classification based on thresholding the burn probabilities. Annual composites of the per-scene outputs are also generated, including (1) maximum burn probability a pixel had across all scenes in the year, (2) the number of scenes in which a pixel was classified as burned, (3) the Julian date of the Landsat scene that a pixel was first classified as burned, and (4) the number of Landsat scenes with unobstructed pixel observations. The algorithm used to produce the BAECV products is a machine-learning approach, trained and evaluated using information about existing burned areas produced by the Monitoring Trends in Burn Severity (MTBS) project. This report describes the provisional BAECV algorithm, and its inputs and outputs.

Introduction

The U.S. Geological Survey (USGS) is developing science-quality, applications-ready, key terrestrial variables and will produce them on an operational basis using historical, current, and future Landsat observations. The terrestrial variables will follow the guidelines established through the Global Climate Observing System and include Climate Data Records (CDRs), which represent geophysical transformations, and Essential Climate Variables (ECVs), which represent specific geophysical and biophysical land properties. CDRs and ECVs offer a framework for producing long-term Landsat datasets suited for monitoring, characterizing, and understanding land-surface change over time.

This document describes the algorithm developed and implemented by the USGS to produce the Burned Area Essential Climate Variable (BAECV) products.

Algorithm description

The BAECV algorithm was designed to automatically extract burned areas from all ecosystems (e.g. forest, shrubland, and grassland) visible in Landsat scenes in the conterminous United States (CONUS). Future versions of the BAECV algorithm will be modified for application in other regions of the world.

Landsat Scene Selection

To train and verify the BAECV algorithm, 29 World Refernce System, version 2 (WRS2) path/rows were selected across the CONUS (Figure 1). Path/row locations were spatially distributed across, what we refer to as algorithm regions, in order to capture major ecosystems and differences in their fire regimes. The algorithm regions were defined by grouping Omernik Level 2 and 3 Ecoregions (Omernik 1987) based on knowledge of fire occurrence patterns and the temporal span over which burned areas are visible in Landsat

scenes. The availability of existing fire information in the Monitoring Trends in Burn Severity (MTBS) database to train and evaluate the algorithm with was also considered when selecting path/rows. Provisional data were produced for several additional path/rows.

The algorithm regions include (1) the Arid West, (2) the Mountain West, (3) the western Great Plains, (4) the eastern Great Plains, and (5) the East. The Arid West was defined by merging the Cold Deserts, Mediterranean California, Upper Gila Mountains, Warm Deserts, and Western Sierra Madre Piedmont Level 2 ecoregions. The Mountain West consists of the Marine West Coast Forests and Western Cordillera Level 2 Ecoregions. The western Great Plains included the West-Central Semi-Arid Prairies Level 2 Ecoregion, and the Central Great Plains, Edwards Plateau, High Plains, Southern Texas Plain/Interior Plains and Hills with Xerophytic Shrub and Oak Forest, and Southwestern Tablelands Level 3 Ecoregions. The eastern Great Plains included the Temperate Prairies Level 2 Ecoregion and the Cross Timbers, Flint Hills, and Texas Blackland Prairies Level 3 Ecoregions. All remaining ecoregions were used to define the East algorithm region.

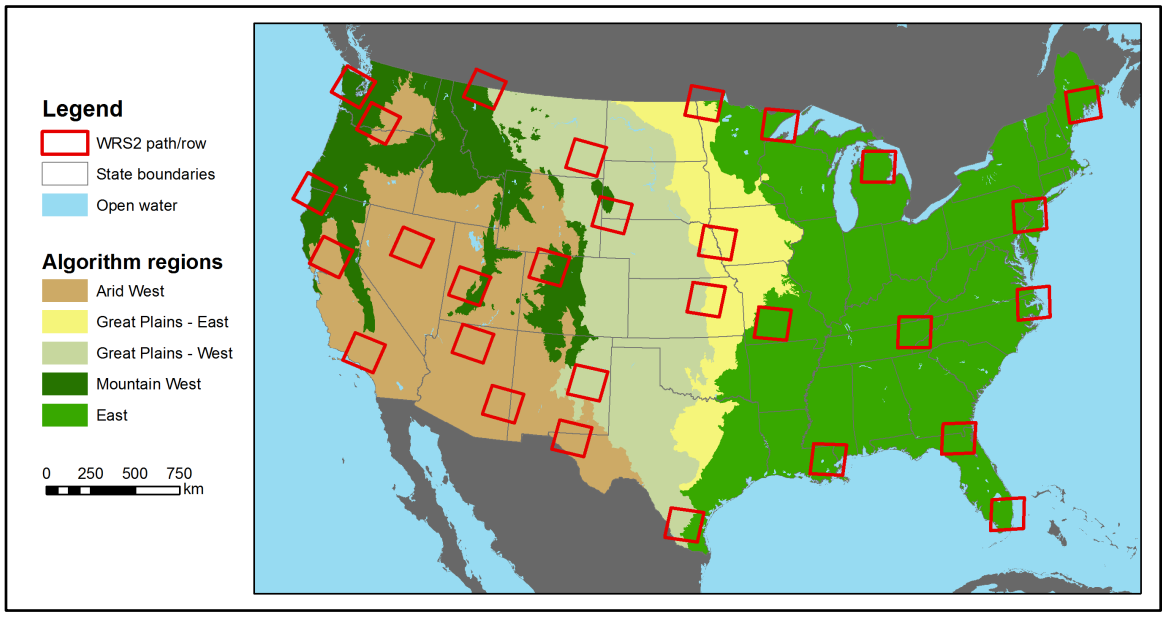

Figure 1. Burned Area Essential Climate Variable algorithm regions and Landsat path/rows used to train and evaluate the algorithm.

For each of the 29 path/rows used for training and verification, we gathered surface reflectance products produced using the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS; Masek and others, 2006) for all available scenes collected by Landsat 4 Thematic Mapper (TM), Landsat 5 TM, Landsat 7 Enhanced Thematic Mapper Plus (ETM+) both with the Scan Line Corrector (SLC) on (1999-2003) and SLC off (2003-present). Scene selection was limited to those with (1) cloud cover less than or equal to 80%, (2) L1T processing, and (3) georeferencing RMSE ≤ 10 m. Both the surface reflectance

and the source metadata were ordered through Earth Science Processing Architecture (ESPA). This resulted in 17,612 Landsat scenes for use in this study (Table 1).

Table 1. Landsat path/rows included for use in training and verification by algorithm region.

			Number of scenes			
Region	Path	Row	Landsat 4	Landsat 5	Landsat 7	Total
East	11	29	2	314	178	494
East	14	32	2	371	188	561
East	14	35	1	392	199	592
East	15	42	3	434	250	687
East	17	39	1	444	227	672
East	19	35	0	386	192	578
East	21	29	2	303	171	476
East	22	39	3	405	223	631
East	25	34	1	374	202	577
East	26	27	0	329	177	506
Eastern Great Plains	28	31	1	331	187	519
Eastern Great Plains	28	33	8	372	213	593
Eastern Great Plains	30	26	1	303	191	495
Western Great Plains	27	41	4	341	217	562
Western Great Plains	32	36	0	451	254	705
Western Great Plains	33	30	3	393	234	630
Western Great Plains	35	28	3	374	214	591
Arid West	32	38	3	473	260	736
Arid West	35	37	8	465	255	728
Arid West	37	35	8	496	264	768
Arid West	38	33	1	430	231	662
Arid West	41	32	0	388	215	603
Arid West	41	36	4	500	262	766
Arid West	44	33	4	442	231	677
Mountain West	35	32	4	430	229	663
Mountain West	41	26	0	355	197	552
Mountain West	45	28	0	379	196	575
Mountain West	46	31	3	379	195	577
Mountain West	47	27	0	300	136	436
Total			70	11,354	6,188	17,612

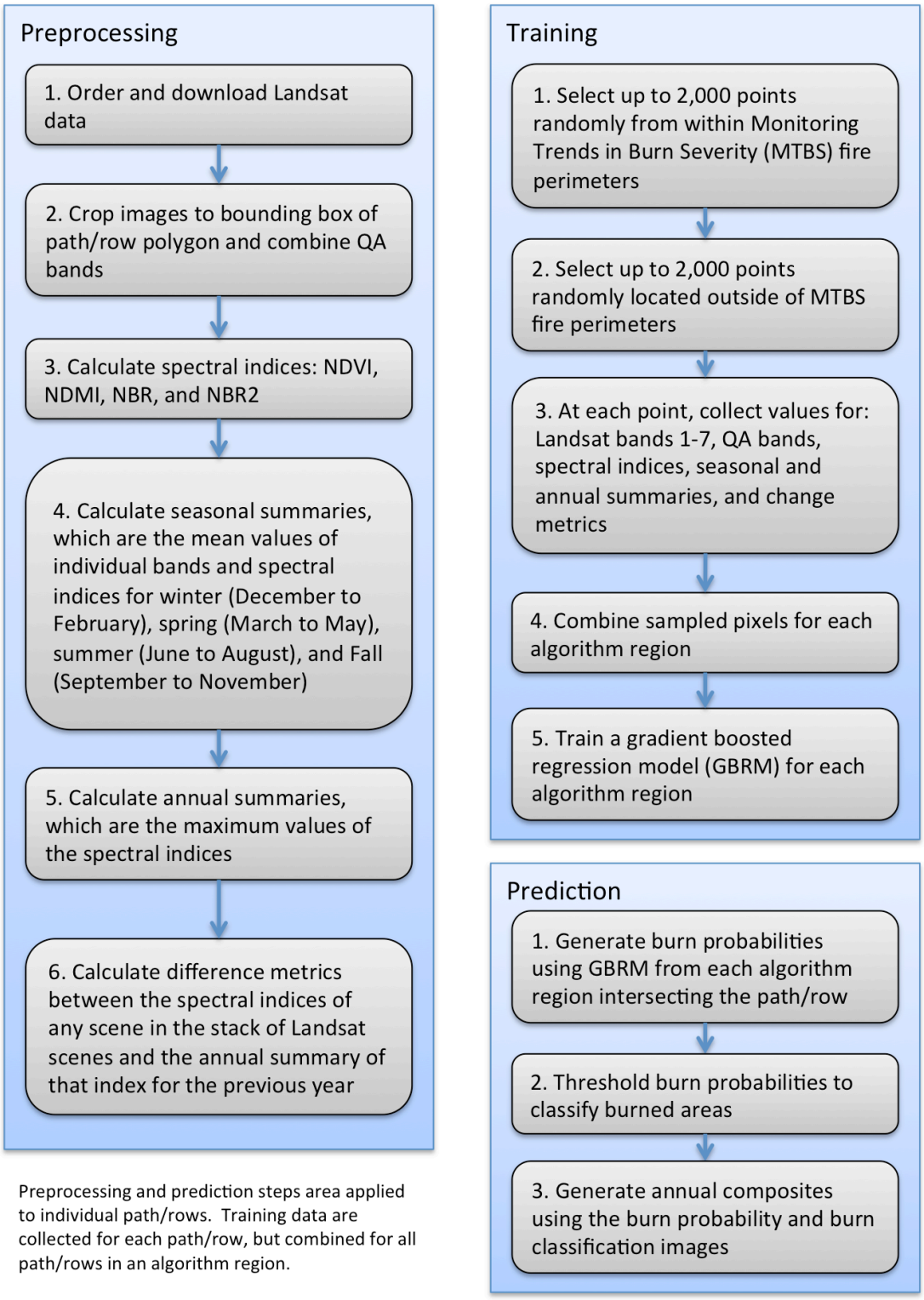

Figure 2. Steps included in preprocessing, training, and prediction for the Burned Area Essential Climate Variable algorithm.

Preprocessing of Landsat scenes

Several preprocessing steps were applied to each Landsat scene prior to training and applying the BAECV algorithm. The spatial coverage of each Landsat scene varied, so after ESPA orders were fulfilled and downloaded, each Landsat scene was cropped to the bounding box of the Landsat path/row polygon. This ensured that each scene had the same number of rows and columns to simplify further processing. During this step, the individual quality assurance (QA) bands were also combined into one single QA band with unique values for each mask (e.g., cloud, water, snow/ice, or fill).

A number of spectral indices were also calculated for each scene. These indices included the normalized difference vegetation index using bands 3 and 4 (NDVI; Tucker, 1979), the normalized moisture difference index using bands 4 and 5 (NDMI; Wilson and Sader, 2002), the normalized burn ratio using bands 4 and 7 (NBR) and a second variant of the normalized burn ratio using bands 5 and 7 (NBR2; Lopez-Garcia and Caselles, 1991; Key and Benson, 1999). These spectral indices are normally floating point values varying between 1.0 and -1.0, but were multiplied by 1000 and represented as integers to reduce memory requirements during processing.

Seasonal and annual summary statistics were derived from the time-series of Landsat scenes in a path/row. Each scene was categorized into a season based on the month the scene was collected; winter included scenes from December of the previous year, and January, and February; spring included scenes from March, April, and May; summer included scenes from June, July, and August; and Fall included scenes from September, October, and November. Seasonal summaries were calculated as the mean for each band and each spectral index. Annual summaries included maximum values for each spectral index (NDVI, NDMI, NBR, NBR2). Pixels flagged as cloudy, water, snow/ice, or fill in the QA masks were excluded when calculating the seasonal and annual summaries. Change metrics for the four spectral indices were also calculated for each scene in the time series of Landsat scenes as the difference between the spectral index of a given scene in the time series and the maximum value of the same spectral index for the previous year.

Burned area training and verification data

The Monitoring Trends in Burn Severity data (MTBS; Eidenshink and others, 2007) were the primary data source used for training and evaluating the results of our algorithm. These data include large fires (≥ 500 acres in the East and ≥ 1000 acres in the West). Each MTBS fire has a fire perimeter shapefile and a categorical burn severity raster layer derived from visual interpretation of pre- and post-fire Landsat imagery. The raster burn severity categories include (1) unburned to very-low severity, (2) low severity, (3) moderate severity, (4) high severity, (5) increased greenness, and (6) masked because of clouds or gaps in Landsat 7 data. We used MTBS severity values of 2, 3, and 4 to indicate that a pixel was burned. The MTBS fires used in this study span the Landsat 4, 5, and 7 epochs (1984 – 2011); and included approximately 17,025 fires burning more than 501,660 square kilometers. For this analysis, the MTBS fire perimeters and severity rasters were cropped to the path/row polygons before additional processing.

Burned Area Essential Climate Variable Algorithm

Burned area probability mapping

The first step in the BAECV approach is to estimate the probability that a pixel was burned. To complete this step, generalized boosted regression models (GBRM) were used. These models tend to produce higher binary classification accuracies than other machine-learning approaches (Hastie and others, 2009). To train and evaluate the GBRM, points were randomly selected within the polygon of each Landsat path and row. Up to 2,000 point locations were located within MTBS perimeters and attributes from each fire were assigned to the points, including the date of burn. An additional 2,000 point locations were randomly located outside of the MTBS perimeters. Points were forced to be a minimum of 30 m apart to avoid sampling the same pixel more than once. At each point location, surface reflectance, thermal bands, and spectral indices values were collected or generated from the individual Landsat scenes. Values for the seasonal summaries, annual summaries, and change metrics at each point were also calculated. Information about the date and sensor (Landsat 4, 5, or 7) of each scene were attached to each point as well. These data were split into training and validation groups based on years; 50% of the years were used for training and 50% were used for validation. Training years included 1984, 1987, 1988, 1991, 1992, 1995, 1997, 1998, 2000, 2001, 2003, 2005, 2006, and 2007. After the training and validation split, all points labeled as burned by the MTBS were retained and an equally sized sample of unburned points were randomly selected.

Using the training point data, a GBRM was trained for each of the 5 regions of the CONUS. Training GBRMs requires users to specify a number of parameters that control the final model structure. The parameters include the (1) number of trees, (2) number of splits per tree, and (3) learning rate between successive trees (Hastie and others, 2009). The number of trees is simply the number of trees to fit in the entire sequence of trees in the GBRM. The second parameter controls the number of splits allowed in each tree. Individual trees are fit in sequence and when fitting the tree, the learning rate specifies the weight to apply to prediction errors from the previous tree in the sequence when fitting a new tree in the sequence.

Through a trial-and-error process, testing the performance of different learning rates and number of splits per tree, the learning rate was set at 0.01 and number of splits per tree at 3. The number of trees used in the GBRMs was selected systematically by evaluating changes in the loss metric for the validation point data as a function of the tree's number in the sequence trees. The objective here was to determine the smallest number of trees in the sequence needed to achieve the minimum value in the loss metric. Initially, GBRMs were fitted for each region using 5,000 trees. The final, reduced number of trees was found by locating the tree at which the change in the loss rate was less than 1% of the moving average of the loss rate from the previous 100 trees. The final number of trees was rounded to the nearest 100 or 1000 (in the direction of lower loss). Area under the curve (AUC) of receiver-operator characteristic (ROC) plots was calculated to judge the accuracy of the final GBRMs (Hanley and McNeil, 1982). After fitting the GBRMs, they were applied to each

image in the time-series of Landsat scenes to generate burn probability images. Pixels flagged as being water, clouds, cloud shadows, snow or ice, and fill in the QA masks were excluded from the analysis.

Burned area classification

The second step in the BAECV algorithm was to threshold the individual burn probability images to produce binary images specifying which pixels were had burned or not. Through ad-hoc visual analysis, the burn probability images often had clumps of pixels with very-high burn probabilities and the clumps were often connected by pixels with lower burn probabilities within MTBS fire. To capture this observed pattern, a region-growing method was implemented for the second step of the BAECV algorithm. First, seeds for potential burned area regions were identified by thresholding the burn probability images into a preliminary binary burned/unburned image; the threshold value is referred to as the 'seed probability threshold'. Second, a seed size threshold was used to remove burned area regions below a minimum size. Third, neighboring pixels were added to seeds if the neighboring pixels had a burn probability greater than or equal to the 'spread probability threshold'. The third step was completed in an iterative fashion until no additional neighboring pixels with burn probabilities above the 'seed probability threshold' could be found. Once the three thresholds were set, binary burn classification images were generated for each Landsat scene in the time series. For the provisional BAECV data, the seed probability threshold was set to 95%, the seed size threshold was set to 45 pixels (10 acres), and the spread probability threshold was set to 85%.

Generation of annual composites, including temporal filtering

In addition to the scene-level burn probability and classification images, a number of annual composite products are generated to assist in the analysis and comparison of the BAECV results. The annual composites include (1) the maximum burn probability across all the individual Landsat scenes in a year, (2) the number of individual Landsat scenes that a pixel was classified as burned, (3) the number of Landsat scenes with unobstructed pixels (pixels that were not clouds, shadows, water, snow/ice, or fill), and (4) the Julian day of the first Landsat scene in which a burned pixel was observed in the year.

Dependencies

The BAECV algorithm depends on input Landsat Surface Reflectance products produced using LEDAPS and the Landsat thermal band. Processing by the BAECV algorithm is dependent upon these data being delivered as either Hierarchical Data Format - Earth Observation System (HDF-EOS) or GeoTIFF image files.

The BAECV algorithm is dependent on annual MTBS burn severity raster layers and burned area polygons for training data.

The prototype code for the BAECV algorithm depends on a number of open-source libraries including: GDAL, numpy, OpenCV, scipy, and sci-kit learn (Table 2).

Table 2. Open source libraries used by the Burned Area Essential Climate Variable algorithm and links to source code and documentation.

Library name	Link
GDAL (Geospatial Data Abstraction Library)	http://www.gdal.org
NumPy	http://www.numpy.org
OpenCV (Open Source Computer Vision)	http://opencv.org
SciPy	http://www.scipy.org
scikit-learn	http://scikit-learn.org/stable/

Inputs

The primary inputs to the BAECV algorithm are the Landsat Surface Reflectance products. These are used with the algorithm both to train and predict. The MTBS burn severity raster layers and burned area polygons (with date information) are required to train the BAECV algorithm.

Outputs

The BAECV outputs include: per-scene outputs corresponding to individual Landsat scenes and annual composites summarizing the individual scenes for a calendar year. Per-scene products include burn probabilities and burn classifications. Burn probability (BP) products are continuous data, representing the probability that a pixel was burned by fire, given what is visible in the Landsat scenes. Burn probabilities range between 0 and 1,000 as they were scaled by a factor of 10. A threshold process was applied to the BP data to produce the burn classification (BC) products. Contiguous burned area patches (identified using a 4-neighbor rule) in the BC data were assigned positive values (patch id), in addition to other attributes quantifying the number of pixels in the seed patch (Area), the number of pixels in the entire patch (FilledArea), and maximum, mean, and minimum burn probability values in the patch (MaxIntensity, MeanIntensity, and MinIntensity attributes, respectively). Both the BP and BC BAECV products retain the QA masks; specified by negative values in the data. Four annual composite BAECV products are provided. The annual composites were generated by stacking all per-scene outputs for a single year. Annual composite outputs include:

1. The maximum value per pixel for burn probability for a year (bp).

2. The count of per-scene burn classification images where the pixel was classified as burned (bc). Values in these imagers are zero or larger. No burned areas were identified in pixels with values of zero in these images.

3. The Julian date of the first per-scene burn classification image where the pixel was classified as burned (bd). Values in these images range between 0 and 366.

4. The number of per-scene images with a cloud-free observation for a given pixel (gc). Values in these images are zero or larger.

Prototype Code

The BAECV algorithm has been implemented in both python and C++ code. This code is available upon request and will be made publicly available after approved by the USGS. Provisional products were produced using the python code.

Verification Methods

The BAECV prototype code, originally written in python, has been used to generate products for a number of path/rows in CONUS. The processing-intensive portions of the python code have been ported to C++ and outputs from both the python and C++ code are similar, but with small differences in burn probabilities because of the small amount of randomness introduced when training GBRMs.

The BAECV outputs are undergoing rigorous validation following the Committee for Earth Observation Land Product Validation protocols. Results of these validation studies will be published as peer-reviewed journal articles as they become available.

Maturity

The BAECV algorithm and products are provisional. Provisional data have not been validated. These data may be subject to significant change and are not citeable until reviewed and approved by the USGS. Additional changes to the algorithm and products are expected pending feedback from provisional data users, to adapt the algorithm to regions beyond CONUS, and to incorporate Landsat 8 scenes.

Acknowledgements

The authors acknowledge the help of Nate Benson, Nicole Brunner, Megan Caldwell, Jeff Eidenshink, Stephen Howard, Carol Mladinich, Kurtis Nelson, Joshua Picotte, Jodi Riegle, and others who provided initial input to the development of the Burned Area Essential Climate Variables. John Dwyer and Calli Jenkerson were also instrumental in providing guidance and support during development. This work would not have been possible without the high-quality data produced from the Monitoring Trends in Burn Severity (MTBS) Project. We also acknowledge the USGS Climate and Land Use Mission Area Land Remote Sensing Program for providing the funding to support this work.

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References Cited

Eidenshink, J., Schwind, B., Brewer, K., Zhu, Z.L., Quayle, B., and Howard, S., 2007, A project for monitoring trends in burn severity: Fire Ecology, v. 3, p. 3-21.

Hanley, J.A., and McNeil, B.J., 1982, The meaning and use of the area under a receiver operating characteristic (ROC) curve: Radiology, v. 143, p. 29-36.

Hastie, T., Tibshirani, R., and Friedman, J., 2009, The Elements of Statistical Learning: Data Mining, Inference, and Prediction: New York, NY, Springer.

Key, C.H., and Benson, N.C., 1999, Measuring and remote sensing of burn severity: the CBI and NBR. In Proceedings Joint Fire Science Conference and Workshop, vol. II, 15–17 June 1999, L.F. Neuenschwander and K.C. Ryan (Eds.), Boise, ID, p. 284 (Boise, ID: University of Idaho and International Association of Wildland Fire).

Lopez-Garcia, M.J., and Caselles, V., 1991, Mapping burns and natural reforestation using Thematic Mapper data: Geocarto International, v. 1, no. 31-37.

Masek, J.G., Vermote, E.F., Saleous, N.E., Wolfe, R., Hall, F.G., Huemmrich, K.F., Gao, F., Kutler, J., and Lim, T.K., 2006, A Landsat surface-reflectance dataset for North America, 1990-2000.: IEEE Geoscience and Remote Sensing Letters, v. 3, no. 1, p. 68-72.

Omernik, J.M., 1987, Ecoregions of the conterminous United-States: Annals of the Association of American Geographers, v. 77, no. 1, p. 118-125.

Tucker, C.J., 1979, Red and Photographic Infrared Linear Combinations for Monitoring Vegetation: Remote Sensing of Environment, v. 8, no. 2, p. 127-150.

Wilson, E.H., and Sader, S.A., 2002, Detection of forest harvest type using multiple dates of Landsat TM imagery: Remote Sensing of Environment, v. 80, p. 385-396.